

GROSS-NET CONVERSION ISSUES

Gianni Betti, Vijay Verma
Università degli Studi di Siena

1. Introduction

The basic requirement in EU-SILC (EU Statistics on Income and Living Conditions) concerning income variables is to record gross income in specified detail at the personal and income component level, but disposable income only as a set of three variable at the total household level. There may be severe practical difficulties for some Member States in collecting income data exactly in this form, whether the data are obtained from registers or directly from respondents in sample surveys. The objective of this paper¹ is to develop, test and recommend procedures on how this problem may be overcome. Both in theory and in practice, this is a complex task. The modelling procedure for net-gross transformation can range from the very complex to the very simple.

At the sophisticated end of the spectrum, tax and social insurance contributions are imputed in as exact a manner as possible using a tax-contribution model, therefore requiring full information on the tax regime operating in a particular country during the reference period. In developing a such sophisticated approach, one possibility is for each Member State to use their own micro-simulation model to provide the output variables required to construct the EU income definition where they are not collected directly from the survey. If such models were to be built from scratch, they would require considerable effort in collating information about the tax regime in each country and repeating it annually. However, many countries have already developed their own models in order to explore the impact of different tax-benefit policies. Such models incorporate household micro-data from nationally representative sources and calculate disposable income for each household under alternative tax/benefit regimes. Although the aim of such models is to simulate the impact of policy changes and so the usual output consists of the micro-level change in household disposable incomes as a result of policy change, they could equally well be used to estimate tax/social insurance contributions under a single tax/benefit regime – as is required for EU-SILC. Euromod provides a well-known example of such modelling in EU-wide context.

Complex models may not be feasible or necessary in all cases. There can be simpler models which take into account only the major elements of the tax regime, and provide sets of ratios or factors which can be applied to individual income components. Alternatively, one may pursue more purely statistical approaches. These are likely to be simpler and more uniform across countries. For instance, net/gross ratios or conversion factors may be determined empirically (statistically) as functions of the household's level and composition of income, household size and composition, and other characteristics available in the survey. These relationships may be determined on the basis of aggregate

¹ Part of the results reported in the paper have been reached in the research carried out within the project “Statistical Services in the Field of EU-SILC, Statistics on Income and Living Conditions. 2001/S 183-125349 LOT 2: Development of Appropriate modelling or imputation to Construct the EU-SILC Target Income Variables for each EU Member States”.

data, such as tax statistics or the national accounts. In principle, information for this purpose may also come from within the income survey itself.

At the other end of the spectrum a single ratio of tax and social insurance contributions paid to gross income may be applied to all income components. This, in the most simple form, was the approach used in ECHP (European Community Household Panel). The basis of the simple approach used in the ECHP was that the (net/gross) ratio can be expressed as a function only of the level of income. Furthermore, empirical data for determining this relationship came from the survey itself – based solely on income from employment, for which both gross and net amounts were collected in the survey. However, it has been clear that the very simple, almost ‘crude’, approach used for ECHP data will not suffice for EU-SILC. For ECHP, in contrast to EU-SILC, the target variables were net (rather than gross) income components. At the same time, income components in the ECHP questionnaire are presumed reported net of tax and other deductions, except for a few components. The most important of these is income from self-employment, which was taken to be reported gross of tax and social insurance payments. Hence, it was not the purpose of the ECHP to provide accurate gross income information. Rather, the objective was to provide a factor which could be used to convert the few components which were reported gross into net values, so that net income, total and by component, of the household could be estimated.

EU-SILC, on the other hand, requires accurate information on both net and gross incomes, and the latter by detailed component and the person level. Hence it is necessary to adopt a more sophisticated approach for EU-SILC, using some more appropriate micro-simulation methodology.

This paper is the first report at an international scientific conference of features and applications of the Siena Micro-Simulation Model (SM2). It will focus on the standardised core of the system and explain how it can handle specific features of diverse fiscal systems and forms in which income of households and persons has been recorded.

The model SM2 is introduced in **Section 2** and is described in the remaining sections by introducing complexities step-by-step. **Section 3** introduces various terms, and describes the basic relationships between them by considering the model in the simplified situation of a person receiving income from a single source and taxed separately as a single-person tax unit. **Section 4** gives a fuller description of the micro-simulation model in the more realistic situation involving more than one income components and multi-person tax units. A number of illustrations (from France, Italy and Spain) are provided. **Section 5** deals with issues arising from diverse forms of the input data input. The form (net, gross, etc.) in which the information has been collected may vary from one individual to another in the same survey. **Section 6** introduces the additional complexity resulting from differences in how particular components of income are treated in the tax regime. A outstanding feature of SM2 is that these special features of the tax system can be captured within the general structure of the model simply by appropriately defining special types of ‘deductions’ and ‘tax credits’ for the component concerned. **Section 7** presents a summary of main results from the micro-simulation system developed for application to country data under EU-SILC. The results presented concern the construction of EU-SILC Target Variables on income in gross and net forms, from the data collected in various forms. So far, it is assumed that data are available for all income components in whatever form and the objective is to convert them to a homogeneous form such as that required for EU-SILC target variables; in the presence of missing data, **Section 8** covers the problem of treating imputation and microsimulation in conjunction. Finally, **Section 9** concludes the paper.

2. The Siena Micro-Simulation Model (SM2)

Income of households is made up of diverse components received by multiple individuals. Its elements may be compiled from different types of sources, which may differ in concepts and definitions and may not refer to exactly the same reference time. The different sources may be subject to differing patterns of response and recording errors, sampling errors, inconsistencies and incompleteness etc. This paper is not concerned with such conceptual and measurement issues, but with the following additional important problem.

Income can be recorded in various forms - such as gross, or net of taxes and/or other retentions at source, or as the final amounts actually received - differently for different components and for different income earners in the population or even in the same household. Aggregating these elements of income into the household's total income and its components, requires not only that information is available on all the elements, but also that it is in a *homogeneous form to permit aggregation*. The form must also be the same for all households to permit aggregation to the population. Furthermore, the same information in *more than one forms* is often required to meet different analytical objectives. Different forms (gross, disposable or net, etc.) are related through complex national tax regimes. This complexity has many aspects. The rules vary by income component and according to characteristics and circumstances of the income recipient.

The Siena Micro-Simulation Model has been developed as a practical tool aimed at providing a robust and convenient procedure for the conversion between net and gross forms of household income. In this paper we describe the logic and standard structure of the SM2. The primary issue which this model has been designed to address may be summarised as follows. Starting from data on household and personal income given in different forms, and on the basis of the prevailing tax regime in a country, the model is designed to estimate full information on income by component, with breakdown of gross amounts into taxes, social insurance contributions of various types, social transfers, and net and disposable income. The immediate context for the development of SM2 has been the requirements of EU-SILC. EU-SILC is a statistical source, developed by European Commission (Eurostat) and implemented by all EU and also other European countries, for the generation of comparable and detailed information on living conditions and income of households and persons. While the source, type and form of input (collected information) varies across and even within countries, the output required at the European level has to be comparable and standardised (Eurostat, 2002).

SM2 is best described as a 'flexible tool' designed to meet these objective in the international, comparative context. An outstanding and unique feature of the SM2 system is that its core consists of a *standardised set of routines* which can handle a great variety of input data forms and national tax systems. Country-specific routines standardise the input data and specify parameters of the national tax system. These constitute the inputs to the central core of the system designed to generate common outputs. The system has been developed to maintain a clear distinction between the standardised and the country-specific parts, and even more importantly, to maximise the part which can be standardised. This feature makes the system an appropriate and convenient tool for multi-country application. The SM2 model has been officially adopted as the recommended procedure for this purpose by the European Commission (Eurostat, 2004).

The central concept in the SM2 procedure is that of 'gross taxable income'. It is connected through rules of the fiscal system to gross income, income as received after tax and other withholdings at source, deductions and tax credits allowed, the tax unit and tax schedule, and finally the net or disposable income. Iterative routines are generally

required to apply these rules. *By specifying deductions and tax credits in appropriate forms, we have been able to impart a common structure to diverse tax regimes and treatments of individual income components* (see Section 6). We consider this feature as an important contribution to micro-simulation work of this type. The paper illustrates these features with reference to some aspects of the French, Italian and Spanish tax systems.

Various micro-simulation models have been developed to simulate taxes, social insurance contributions, benefits and other transfers received to affect the transformation between gross and net forms of income, mediated through complexities of the national fiscal systems. Important examples are Euromod (2001), and a host of similar national micro-simulation models. The objective and orientation in developing SM2 has been somewhat different (though it shares much in objectives and methods with existing micro-simulation models). A system such as Euromod may be viewed as a ‘comprehensive facility’, while SM2 aims to be a ‘flexible tool’. On the basis of specific data-sets incorporated into the system, Euromod provides a facility for micro-simulation of the effect of varying parameters of the tax-benefit system. Simulation of benefits under different regimes has therefore to be an essential part of the system. By contrast, SM2 is a tool to be applied to diverse data sets to generate a required set of income-related target or analysis variables in a standardised form. It is fully ‘data based’ and presently does not incorporate simulation of benefits or any other income components: it is taken as given that information on all income components has been collected, compiled or imputed in some form, and that the objective is to convert it to the standard form under a given national tax system.² The emphasis is on flexibility, so as to be able to deal with an annual flux of data in different forms across and within national surveys, and also with periodic changes in the national tax systems.

3. Terminology: forms of income and their relationship

The relationship among different forms of income is summarised in Table 1.

Gross income (GG, G) of an individual, household or other tax unit is the total income from all sources received during a reference period, before any deductions for tax or social insurance contribution. In the model it is useful to distinguish between gross income (GG) including employer's social insurance contributions, and the gross (G) including only other social insurance contributions. The social insurance contributions are generally a function of G, rather than of GG.

Social insurance contributions. Normally, these contributions apply only to income from work and include (i) employer's contribution on behalf of persons in employment (SS); and (ii) employee's and self-employed person's contributions (S). In most national systems, the social insurance contributions are *component-specific*, determined independently of other components of income. However, the system can be more elaborate. For instance, in France practically all income components are subject to social insurance contributions. Contributions for some components are a function of the *combined income* from a set of several components. Furthermore, a part of the social insurance contributions are themselves subject to income tax. Nevertheless, these complications merely make the algorithm specifying the various functional relationships

² See Section 8 concerning issues arising in relation to missing data and imputation.

more elaborate, but there is no problem in handling them within the common structure of the SM2 model.

Tax unit. This refers to the set of individuals whose incomes are pooled together for the purpose of determining tax and social insurance liability. Normally it refers to related individuals from the same household, and often to one individual.

Gross taxable income (H) is gross income less social insurance contributions: $H = G - S$.

Deductions (D) refers to part of gross taxable income which is exempt from tax.

Net taxable income (Y) is obtained by subtracting from gross taxable income 'deductions', i.e. the part which is tax exempt: $Y = H - D$.



Tax due (W). Initial tax due is computed as a function of net taxable income, $W = W(Y)$. This is determined by the prevailing income tax schedule. For the main part, this normally involves *pooling of income across components and individuals in the tax unit* to which the common tax schedule applies.

Tax credits (C). The tax liability is normally reduced by tax credits.

Tax paid (X). Deduction of these tax credits from the tax due gives the final tax to be paid: $X = W - C$.

Total net income (N) is total gross taxable income less tax paid: $N = H - X$.

Table 1 - Basic relationship among forms of income (one person, single source)

	form	relationship	comment
1	Gross income	G	GG = G + SS(G)
2	Social insurance contributions	S = S(G)	
3	Gross taxable income	H = G - S	
4		tax and other deductions at source	XS = H XST = H - T(H) XT = H + S(G) - T(H)
5	Deductions	D = D(H)	
6	Net taxable income	Y = H - D	
7	Tax due	W = W(Y)	
8	Tax credits	C = C(Y)	
9	Tax paid	X = W - C	
10	Net income	N = H - X	

SS: employer's social insurance contribution. Retentions at source: XS=social insurance only; XST=tax and social insurance; XT=tax only.

In certain systems, income as initially received is subject to retention at source of tax and/or social insurance contributions T(H). For instance, for income subject to both these retentions we have: $XTS = G - S(G) - T(H) = H - T(H)$. Unlike the amount of 'final tax due'

W or X which is determined through complex rules involving pooled income over individuals and components, the relationship $T(H)$ determining tax retention at source is often component-specific and much simpler.³ The same applies to any income components which are taxed separately from pooled income as defined above (such as at a flat or some other component-specific rate).

Gross income, social insurance contributions, gross taxable income, tax and social insurance withheld at source, income received after retentions at source, and some deductions and tax credits are defined at the level of individual income components. Other quantities may involve aggregation over components (and also over individuals in a tax unit). It is useful to first look at the basic relationships between different quantities for the simple case of a single-person tax unit, receiving income only from one source. Table 1 summarises these.

4. Gross-to-Net conversion algorithm

Figure 1 shows the basic relationship between gross and net forms of income in the more realistic situation when more than one income components and possibly more than one individuals in the tax unit are involved. Even in this case, the relationships between gross taxable income for a particular component, H_i , and quantities like gross income G_i and income after retentions at source XST_i are generally simple, dependent only on the concerned income component i , and determined independently of other components and other persons in the tax unit. The same applies to the relationship between H_i and net N_i for components which are taxed separately at a flat rate or a rate determined only by the level of income from that component, and of course also for tax exempt components. Sometimes, dependence of the relationship on other sources of income may also be involved, but mostly these are simply in the form of upper limits which may apply to certain quantities pooled over more than one component.

In practice, all or most taxable income is pooled together over components and over persons in the tax units for the purpose of determining the amount of tax due.

The relationship between H_i and N_i for components in the pool is, therefore, more complex. Going from H_i to N_i involves less complexity since the relationships (the tax rules) are a function of the known H_i . These relationships are specified in more detail in Table 2. Going from given N_i to H_i required more complex iterative solutions, and are described in the Section 5.

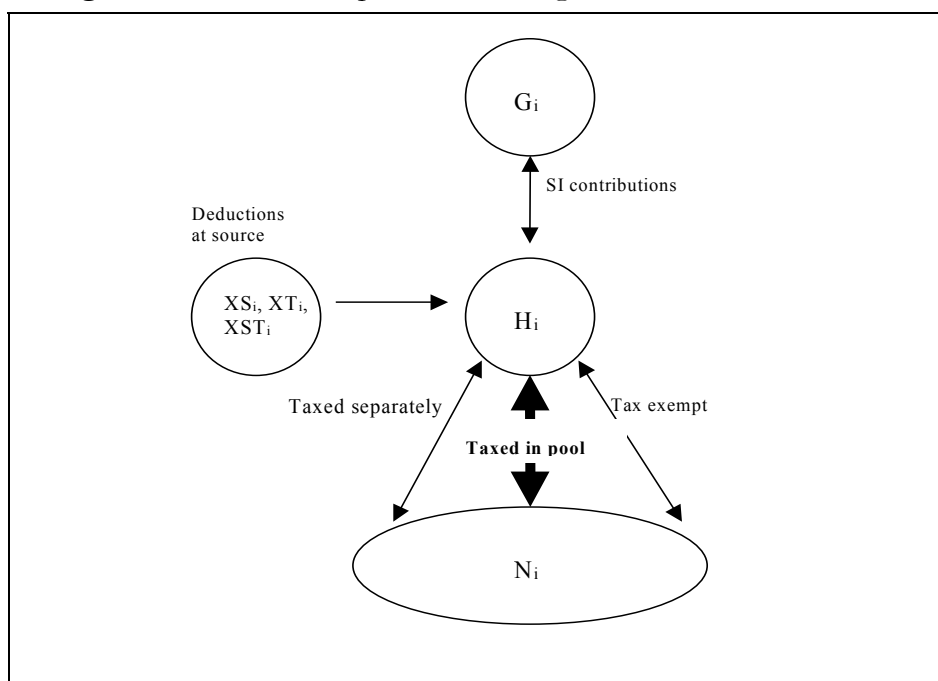
Social insurance contributions

The social insurance contributions S_i , if applicable to the component, are generally a function of the gross amount G_i (which in the case of employment income is defined to exclude the employer's social insurance contribution). However, some more complex situations can be allowed for in the model while retaining its basic structure. Specifically, it can allow for the dependence of S_i for any particular component i on any set of income components, i.e., a functional relationship of the form $S_i = S_i(G_I)$, where subscript I refers to any set of income components (normally including the particular i , of course). In the French system for instance, the pooled contributions for a number of components may be

³ However, sometimes the amount retained may be determined by individual arrangements rather than on the basis of fixed rules of the fiscal system, in which case the relationship $T(H)$ has to be determined at the micro-level.

subject to a common maximum limit. The functional relationship $S_i(G_i)$ is specific to the component and the country. This is specified (and ‘called’ as a subroutine in the SM2 application programs) separately from the common structure represented in Table 2.

Figure 1 - Basic relationship between net and gross amounts



Deductions

(Net) taxable income (row 7) is obtained by subtracting from gross taxable income the part which is tax exempt (‘deductions’). These deductions are a certain function of gross taxable income. These may be of two types: (i) specific deductions which apply to the particular income components D_i (row 4); and (ii) common deductions which apply to the (remaining taxable) income from all sources together (row 6).

In case (i), component-specific deductions, in most situations the functional relationship $D_i(H_i)$ is specific to the component i , i.e., D_i depends only on the gross taxable income H_i for the component concerned. As a generalisation, the model can allow for the dependence of D_i for any particular component i on any set of income components, i.e., a functional relationship of the form $D_i = D_i(H_I)$, - or even more generally as $D_i = D_i(H_I, G_I)$ - where subscript I refers to any set of income components (normally including the particular i).

In case (ii), common deductions, a functional relationship of the form $D_0(H)$ is in terms of total gross taxable income i.e. all components put together. Both types of functions are of course country-specific. Again, these relationships can be specified separately from the common structure represented in the table.

Aggregation

After the removal of component-specific deductions, it is necessary to pool the income across components which are treated together for taxation purposes and over individuals in the same tax unit. Certain income components may be excluded from this common pool and taxed separately; this type of situation is accommodated in the model (see Section 6).

Table 2 - Gross-to-Net conversion algorithm

	Income measure	total	by component⁽¹⁾
1	GROSS ⁽²⁾	$G = \sum G_i \leftarrow$	G_i
2	Social Insurance contribution		$S_i = S_i(G_i)$
3	GROSS TAXABLE	$H = \sum H_i \leftarrow$	$H_i = G_i - S_i$
4	Component-specific deductions		$D_i = D_i(H_i)$
Aggregation over components and individuals in tax unit			
5	TAXABLE INCOME	$Y = \sum Y_i \leftarrow$	$Y_i = H_i - D_i$
6	Common deductions	$D_0 = D_0(H)$	
7	Taxable income(0)	$Y_0 = Y - D_0$	
8	Tax due(0)	$W_0 = W_0(Y_0)$	
9	Common tax credits	$C_0 = C_0(Y_0)$	
10	TAX DUE	$W = W_0 - C_0$	
11	Component-specific tax credits	$C = \sum C_i \leftarrow$	$C_i = C_i(Y_i)$
12	TAX PAID	$X = W - C$	
13	TOTAL NET	$N = H - X$	
14	Tax rate(0)	$R_0 = X/H$	
15	TAX RATE = TAX DUE/ TAXABLE INCOME	$R = W/Y$	
Disaggregation – personal income by component			
16	Proportionate tax by component		$X_i = R * Y_i - C_i$
17	NET BY COMPONENT		$N_i = H_i - X_i$

(1) The functional relationships in this column may be somewhat more complex or varied.

(2) Gross including employers' social insurance contribution (SS) is: $GG = G + SS(G_i)$

Tax due and tax credits

Initial tax due is computed as a function of total taxable income (row 8). This is determined by the country's 'basic' income tax schedule, normally applied to pooled income from different sources. This tax liability is normally reduced by tax credits. Tax credits are mostly based on characteristics of the unit (single parent, pensioner, etc.) or are given in compensation for particular expenses (medical, educational, etc.), i.e., are not specific to a particular income *source*. We refer to these as 'common tax credits' (row 9); these are normally expressed as a function of the total taxable income. The result is a more precise expression of 'total tax due' (row 10). In addition to the common tax credits, there may also be component-specific tax credits (row 11). Generally, these are based on net taxable income for the component concerned. However, the functional relationship may be more complex involving other components of income and/or income in other forms (gross, gross taxable, etc.), as noted above.

Tax paid and net income

Deduction of these tax credits from the tax due (as defined in row 10), gives the final tax

to be paid (row 12): total tax to be actually paid is tax due, less all (common as well as component-specific) tax credits. Total net income is total gross taxable income less tax paid (row 13).⁴ The above two quantities, tax paid and net income (rows 12-13) refer at this stage to total income, and not to income by individual components.

Tax rate

This refers to the effective tax rate which applies to pooled components. Tax rate in Table 2 has been defined in two forms. (i) The first (row 14) is a descriptive measure (R_0). It is the ratio of the total amount of tax to be paid, to the total gross taxable income (row 12/row 3). Hence it is indicative of the overall tax burden. (ii) The second (row 15) provides a more analytical measure R in the following sense. It is the ratio of the total amount of tax due before taking into account any *component-specific tax credits* (row 11), to the total taxable income after removing *component-specific deduction* (row 4). By removing all known component-specific aspects, that is component-specific deductions and tax credits, it can be seen as the *common rate* which applies to all taxable income, from whatever source, which has been pooled and subject to a common tax schedule.

Parameter R has two functions. Firstly, it provides a means for the disaggregation of total tax and net income into component when required (see below). Secondly, it is the parameter of the iteration in going from net to gross, as described in the next section. Its role is even more important in the presence of missing data where modelling has to be used in conjunction with imputation (Section 8). We have explored these issues further in a separate paper (Betti *et al.*, 2003).

Disaggregation of tax and net income by component

This common 'tax rate' R can be seen, without any added assumptions, as a rate applying to each component individually, and not merely some average rate applicable only at the level of total income. This permits the decomposition of tax paid by income components (row 16), and consequently the decomposition of total net income into components (row 17). This decomposition is essential for the construction of variables such as net income before and after social transfers which are required in EU-SILC. For research and policy purposes, decomposition of net income is usually required in less detail than the breakdown of gross income. In any case, this sort of breakdown does not affect the performance of the rest of the system in the model in any way.

Country-specific schedules

The last two columns of Table 2 define the various income measures in terms of measures defined in the preceding rows; those in the first column concern total income, and in the second concern income components. The table involves six country-specific relationships or tax schedules:

-three concerning total income	$D_0 = D_0(H), W_0 = W_0(Y_0), C_0 = C_0(Y_0)$
-another 3 specific to each component (i)	$S_i = S_i(G_i), D_i = D_i(H_i), C_i = C_i(Y_i).$
If applicable, there may be parameters determining tax retention at source, $T_i = T_i(H_i)$, and taxable part of social insurance contributions, ΔS_i etc.	

⁴ Strictly, this may be referred to as 'disposable income'. Sometimes the term 'net income' is used for gross income less social insurance deductions and tax due, while the concept of disposable income also takes into account inter-household and some other transfers.

The functional dependence can be somewhat more complex than indicated above, as explained earlier. In addition, there may be parameters determining retentions at source, taxation of parts of social insurance contributions, etc. Finally, it should be mentioned that the application of various formulae and relationships requires certain constraints to be met, such as to ensure that all quantities which, to be meaningful, must be non-negative are in fact so. It is not useful to list here such (and many other) programming details.

5. The core iterative procedure

The form in which data on income by component are available may vary from one country (tax regime) to another, and also among individuals and households within the same country. There are two dimensions of variation:

A. Whether or not a particular component is subject to social insurance contributions and to income tax. Income tax and other deductions may apply in various forms. The basic distinction is between: (i) income which for the purpose of taxation is pooled together, across components and also across individuals in some appropriately defined tax unit; and (ii) other incomes each treated separately. In either case, information for the appropriate treatment of each component can be generally compiled at the aggregate level and need not be collected at the micro level.

B. The form in which the information has been collected (such as net or gross). This may generally vary from one individual to another in the same survey, though a uniform reporting form may prevail for some components. In any case, the information on the form in which the data are available is required at the micro-level.

We first describe the standardised ‘core’ of the SM2 system, taking account of complexities resulting from the information reported in diverse forms but assuming for the moment that all income components over individuals in the tax unit are pooled together and subject to a common tax schedule.

Table 3 - Calculation of H_i according to the form of data specification

Set H given value $X_i =$	XS_i	$H_i = XS_i$	
	G_i	$H_i = G_i - S_i(G_i)$	
	XT_i	$H_i = G_i - S_i(G_i)$ where $G_i = XT_i + T_i(H_i)$	Simple iteration, generally separately for each component
	XTS_i	$H_i = XTS_i + T_i(H_i)$	
Set N given value $X_i =$	N_i	$H_i = Y_i + D_i(H_i)$, where $Y_i = [H_i - N_i + C_i(Y_i)] / R$	Double iteration: (i) with assumed R, for each component in turn, and (ii) for determining R, common to all pooled components

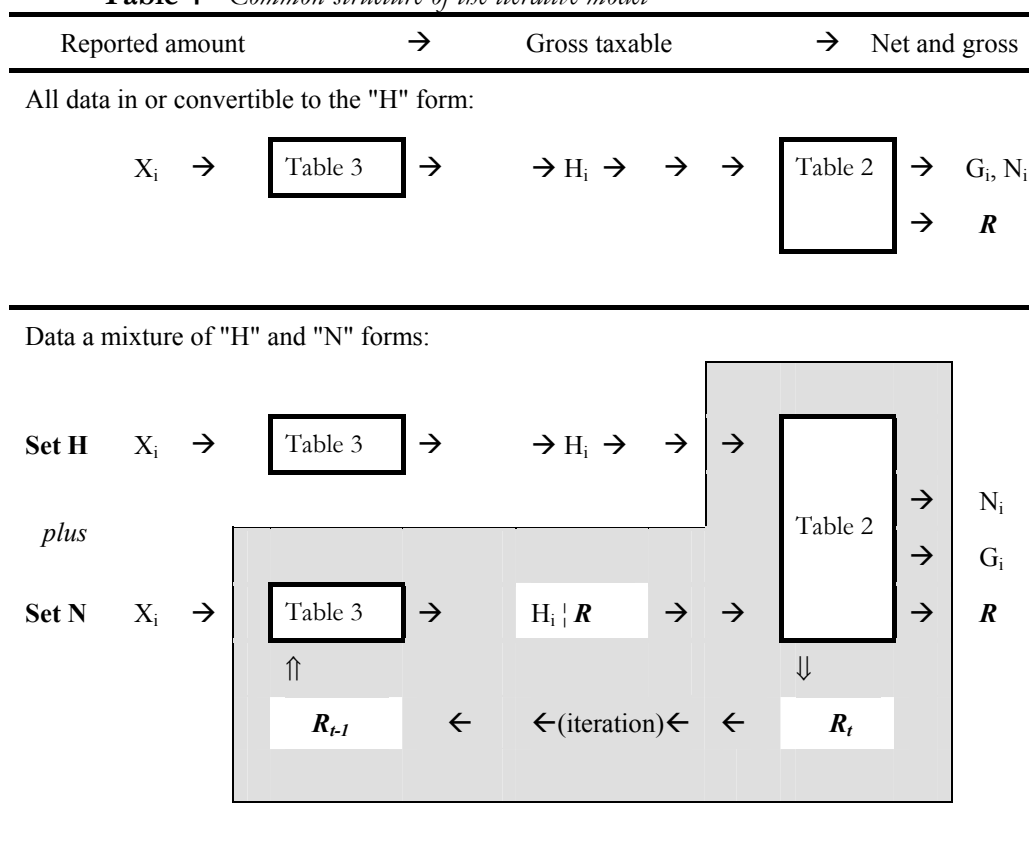
Table 3 shows the procedure for converting the reported amount with any combinations of the above dimensions of variation into a standard form. The top part of the table (‘set H’) refers to the simpler situation in which the input information is in any form other than ‘final net’ N_i (see also Table 4). In this case, it is convenient to take ‘gross taxable income’ H_i as the standard: $X_i = [G_i, H_i, XS_i, XTS_i, XT_i] \Rightarrow H_i$.

Here X_i is the one of the forms G_i, \dots, XT_i in which the information has been collected, H_i is the required gross taxable income, and the other quantities refer to income received after tax and/or social insurance contributions deducted at source.

We take ‘retention at source’ to mean that the amount of tax and/or social insurance contribution has been assessed depending only on the income received from the particular source concerned by the individual concerned, not taking into account the income received from any other sources or the individual's personal characteristics. We take this conversion to involve the component and country-specific functional relationships or schedules, namely $S_i = S_i(G_i)$, social insurance contributions, and $T_i = T_i(H_i)$, tax retention at source.⁵

In a majority of the cases, H_i can be determined directly from the collected amount, for instance from gross amount (G_i) reported for an income component i subject to social insurance contributions: $H_i = G_i - S_i(G_i)$. In other cases, an iterative procedure may be required. However, generally the iteration is very simple and converges quickly. This is because by and large component-specific schedules apply to each component separately. There are no other parameters to be estimated. The need for numerical iteration arises simply from the fact that the unknown quantity to be determined (H_i) appears in an implicit equation.

Table 4 - Common structure of the iterative model



Set of variables N: set of income components which are subject to income tax (irrespective of whether the component is also subject to social insurance contributions), and for which the ‘final net’ amount ($X_i=N_i$) has been specified in the data collected.

Set of variables H: all other income component (not subject to tax, or for which the data has been collected in a form other than the ‘final net’ amount).

⁵ Sometimes in practice, however, retentions at source may not be according to fixed schedules, but according to arrangements determined at the individual level. In such cases the common rules do not apply and the information must be collected and used at the micro level.

The situation is more complicated when the reported amount is in the form ‘final net’ N_i (‘set N’ in Tables 3 and 4). Going from N_i to H_i in fact involves a double iterative loop. The inner loop of iteration is applied with an assumed value of the parameter ‘tax rate’ (R , as defined in Table 2). Once this has been done for every income component in the group (including over all individuals in the same tax unit), an outer iterative loop obtains a convergent value of parameter R which is common to all those components. The N_i to H_i conversion process is therefore considerably more complex. Furthermore, this complexity is substantially increased in the presence of missing data, where the modelling and imputation procedures will have to be applied interactively in conjunction (Betti *et al.*, 2003).

Table 4 demonstrates the common structure of the iterative procedure for the general case where different reporting forms may exist simultaneously in the data for different components, individuals and tax units. As noted, the income components may be divided into two sets, ‘N’ and ‘H’, depending on whether the amount reported is ‘final net’ (N_i), or is in some other form $X_i = (G_i, X_{Si}, X_{Ti}, X_{TSi}, H_i)$ more directly convertible to the gross taxable form H_i . The procedure may be applied as follows. The required H_i quantities for set H are computed (only once) using Table 3, and form an input into the iterative cycle for parameter R required for set N. The parameter is best estimated by using information on all income components from both the sets N and H.

6. Special deductions and tax credits: a device to treat diversity

A remarkable feature of SM2 is that by appropriately defining certain ‘special’ deductions and tax credits, many special features and complexities of different tax regimes can be incorporated into the standardised procedures described in the previous section without altering them in any way.

Deductions refer to the part of gross taxable income which is tax exempt. As noted, these deductions may be *component-specific*, or may be *common deductions* which apply to taxable income as a whole to yield net taxable income. Initial tax due is computed as a function of total net taxable income. This tax liability is normally reduced by tax credits. Again, these may be component-specific, or may be common credits which apply to the initial tax due as a whole. In addition to these ‘normal’ deductions and tax credits, *we can define ‘special’ component-specific deductions and tax credits to accommodate variations in the form in which the component is taxed without altering any other aspect of model specification.*

Table 5 lists a number of such possibilities. In fact, it covers all the situations we encountered in application of the SM2 to data from the European Community Household Panel for France, Italy and Spain to construct a standardised set of gross and net income variables for households and individuals (see Section 7).

Consider for instance the common situation with one component (such as family benefits) tax exempt, and the remaining components pooled together and subject to a common tax regime (row 1 of Table 5). By simply specifying ‘special deduction’ for the tax exempt component as $D_i=H_i$, i.e. the same as its gross taxable amount, we automatically retain its tax-exempt nature and it is no more necessary to separate it from rest of the pool. It makes no contribution to the total net taxable income, and its original gross taxable income appears automatically as a part of the final net income. Similarly, if a component is taxed at a flat rate (say f_i) separately from the pool (row 2), we can simply specify its ‘special deduction’ as $D_i=H_i$, and its ‘special tax credit’ as a *negative* quantity $C_i=-f_i*H_i$. It makes

no contribution to the tax liability of the pool, but the final tax liability is automatically increased by the appropriate amount. Again, no other treatment separate from the pool is required for this component.

The situation in the case of a component tax-exempt at a flat rate is just the opposite (row 3): it is fully retained in the pool of taxable income, and the flat rate tax exemption accommodated as a (positive) tax credit. Deductions for expenses can be specified as common deductions applicable to the total income i.e. not associated with any specific income component (row 4), and similarly for tax credits for expenses (row 5). Sometimes components are subject to ‘double’ taxation. For example, in Italy self-employment income is liable to income tax as a part of the total income in the normal way, and also to an additional tax (‘IRAP’) at a rate depending only on the component concerned. This complexity is easily handled as shown in row 7: the component is fully retained in the taxable income pool, and the flat rate double taxation on it is accommodated as a *negative* tax credit.

The last case (row 8) is an important one, as it handles a special and complicating factor in the treatment of social insurance contributions, which themselves are subject to tax, as in France for instance. By specifying the taxable part of social insurance contributions as *negative deductions* from (i.e. in effect as additions to) gross taxable income ($H_i = G_i - S_i$), the net taxable income (the amount actually subject to tax) is augmented by the taxable part of the social insurance contribution, say ΔS_i : $Y_i = H_i + \Delta S_i$. No further special treatment of this feature of the system is required in the model.

Table 5 - Examples of special deductions and tax credits

	Form of taxation of component i	Special deduction	Special tax credit
1	Tax exempt	$D_i = H_i$	-
2	Taxed at flat rate f_i	$D_i = H_i$	$C_i = -f_i * H_i$
3	Tax-exempt at flat rate f_i	-	$C_i = +f_i * H_i$
4	Deductions for expenses	+common deductions	-
5	Tax credit for expenses	-	+common tax credits
6	Special tax not related to income	-	-common tax credits
7	Double taxation at flat rate f_i	-	$C_i = -f_i * H_i$
8	Part ΔS_i of social insurance contributions subject to tax	$-\Delta S_i$	-

Different forms may apply to cases like 3,4 and 7: for instance the tax rate being a more general function of the amount of income involved for the component concerned.

7. Applications of SM2 to France, Italy and Spain

This section presents a summary of main results from the micro-simulation system SM2 developed in particular for application to country data under EU-SILC. The results presented concern the construction of EU-SILC Target Variables on income in gross and net forms, from the data collected in various forms. Detailed applications have been developed for France, Italy and Spain with ECHP data as the input.

Data sources and reference years for the three countries to which the model has been applied in detail are as follows:

country	Data source and reference year	Fiscal system reference year	Main form of input data from the survey (by component)	
France	ECHP1998	1998	Gross of tax but net of social insurance deductions	XS
Italy	ECHP1998 <i>Supplementary:</i> ISTAT household budget survey Bank of Italy income survey	2003	Final net income	N
Spain	ECHP1999	1999 (applied) 2003 (also studied)	Net of taxation and social insurance deducted at source	XTS

The data used were for the following reference years: 1998 (i.e., ECHP Wave 6, survey conducted during 1999) for France and Italy, and 1999 (ECHP wave 7) for Spain. The primary interest in the following presentation of empirical results is methodological rather than substantive. The reader should bear in mind that both the data and the tax systems used refer to the situation a number of years ago, and may not reflect the current situation.

The choice of the reference year for the fiscal system had two conflicting requirements: (i) it should be the same as the income reference year for consistency; but ideally (ii) it should reflect the situation when EU-SILC began (reference year 2003), so as to be directly applicable to EU-SILC. The second requirement is important if there has been significant changes in the fiscal system since the data reference year.

In France, we gave precedence to requirement (i) concerning consistency, and applied the 1998 fiscal system to survey data for the same year. The application remains useful for EU-SILC in so far as have not been significant changes to the fiscal system since that time. In the absence of such changes, it would be a relatively straightforward matter to update the fiscal parameters to reflect the most recent situation in application to EU-SILC data.

By contrast, in Italy there have been significant changes in the fiscal system since the income reference year (1998) for which the most recent ECHP data are available. To be useful for future EU-SILC applications, it was necessary to take into account the fiscal rules at the start of EU-SILC. Hence we have given precedence to requirement (ii). To 1998 data (adjusted only by a single factor to reflect changes in the overall income levels between 1998 and 2003), the 2003 fiscal rules were applied. This inconsistency can be expected to introduce some distortions in the results, but is preferred because the model is more relevant for actual application to EU-SILC.

The ECHP data in Italy are supplemented from alternative sources to model certain deductions and tax credits which depend on the level of consumption.

In the case of Spain, there were major changes between 1998 and 1999 in the fiscal system, and therefore data for the later year were used for these illustrations. We have applied the 1999 fiscal rules to data for the same year. The application remains important for EU-SILC since there have been no major changes in the system in the interviewing years.

The forms in which input (ECHP) data are available are different between the three countries, and hence provide a good range for illustrating and testing the model. With the possible exception of a few components, the following applies.

- In France, ECHP data on income by component were collected gross of income tax but net of social insurance contributions.⁶
- In Italy, the ECHP data are taken to be in the ‘final net’ form, i.e., after the deduction of all taxes and social insurance contributions due. The main exception is income from self-employment, which was reported gross of tax and social insurance contributions.⁷
- In the available data for Spain, the reported income components are net of tax and social insurance deductions at source.⁸ The main exception is again income from self-employment.

It is important to point out that application of the model requires some data in more detail (or in a different form) than available in ECHP Users’ Data Base (UDB). We are grateful to Eurostat and the countries concerned for making available data from the original Production data Base (PDB).

7.1 Applications of SM2 to France

Table 6 shows comparison of SM2 application to France with figures published by INSEE. *The agreement at the overall level, both in terms of mean income, and more importantly, in terms of the structure of revenu avant imports in terms of social insurance contributions, taxes and net (disposable) income is, in our view, excellent.* Comparison for other parts of the social insurance contributions not included in *revenu avant imports* is not possible from these data. We may also note a point of detail in the comparison in Table 6: information is not available in the ECHP data set to estimate household tax, and this component is confounded with estimated ‘net’ from our model SM2. To facilitate comparison, we have simply taken the average value (1.1%) from INSEE results.

Table 6 – *Application to France: comparison with INSEE*

	Composition of gross income (INSEE)		comparison:	
	mean amount	distribution	SM2	INSEE
[1] net+household tax	67,858	67.2		
[1a] net		66.4	86.6	85.7
[1b] household tax		0.8	1.1	1.1
[2] personal income tax	4,086	4.0	5.3	5.9
[3] CSG,CRDS, SI on capital income	5,431	5.4	7.0	7.3
subtotal [1]+[2]+[3]	77,375	76.7	100.0	100.0
[4] other personal SI contributions	6,180	6.1	Mean/household (Euro)	
[5] employer’s SI contribution	17,397	17.2	30,200	29,856
total social insurance contributions	29,008	28.7		
total gross	100,952	100.0		

Source: INSEE: INSEE PREMIERE, n. 916, August 2003, ‘Des ménages modesties aux ménages aisés: des sources de revenus différentes’. SM2: France ECHP Wave 6.

⁶ The reported income data are also net of that part of social insurance contributions which are themselves subject to income tax. This is a unique feature of the French system.

⁷ This in fact is the form in which ECHP data in most countries is believed to have been reported.

⁸ Note that this may (and often does) differ from the ‘final net’, i.e., after the deduction of all taxes and social insurance contributions due.

Table 7 – France EU-SILC target variables: distribution of income by component

	mean amount		ratio	% distribution	
	gross	net	net/gross	gross	net
	(1)	(2)	(3)	(4)	(5)
income from work	74,269	44,121	59.4	75.0	66.3
PY010 employee cash or near cash income	41,228	38,951	94.5	41.6	58.5
employer's SI contribution	17,059			17.2	
employee's SI contribution	9,549			9.6	
PY050 cash benefits or losses from self-employment	5,576	5,170	92.7	5.6	7.8
Self-employed SI contribution	857			0.9	
property income	2,664	2,217	83.2	2.7	3.3
HY090 interest, dividends, profit from capital	1,704	1,531	89.9	1.7	2.3
capital income SI contribution	185			0.2	
HY040 income from rental of a property or land	775	685	88.5	0.8	1.0
taxable benefits	20,383	18,526	90.9	20.6	27.8
PY090 unemployment benefits	1,693	1,620	95.7	1.7	2.4
unemployed SI contribution	95			0.1	
PY100 old-age benefits	14,170	13,278	93.7	14.3	20.0
PY110 survivor benefits	897	822	91.6	0.9	1.2
pension SI contribution	681			0.7	
PY120 sickness benefits	669	669	100.0	0.7	1.0
PY130 disability benefits	394	372	94.3	0.4	0.6
disability SI contribution	18			0.0	
HY050 family related allowances	1,741	1,741	100.0	1.8	2.6
PY150 other personal benefits	25	25	100.0	0.0	0.0
tax-exempt social transfers	1,674	1,674	100.0	1.7	2.5
PY140 education-related allowances	152	152	100.0	0.2	0.2
HY060 social assistance	182	182	100.0	0.2	0.3
HY070 housing allowances	720	720	100.0	0.7	1.1
HY080 regular inter-household cash transfer received	621	621	100.0	0.6	0.9
total	98,990	66,539	67.2	100.0	100.0

French Francs 1998, per capita

Table 7 shows the distribution of estimated gross income by component. These are the main EU-SILC target variables: all EU-SILC income target variables can be constructed on the basis of appropriate aggregation of such classification by component. The model also provides the same breakdown for net income and the net-to-gross ratio, though these are not included in the required target variables in EU-SILC. Because of differences in component-specific deductions and tax credits, and also in the social insurance contributions, the net/gross ratio varies by component. The net-to-gross ratio is much lower for income from work (around 60%) than for other components. This results from the social insurance contributions to which such income is subject (in France, many other types of income is also subject to SI contributions, but at much lower rates). Leaving aside the effect of social insurance contributions, the ratio of *net* to *gross taxable income* varies approximately from the low of 89% for property income, to 94% for work income, 94% for various taxable benefits, to of course 100% for housing, social assistance and other tax-exempt benefits. These results appear plausible, though only limited external data are at hand to validate the breakdown in detail by component.

7.2 Applications of SM2 to Italy

Table 8 shows comparison of SM2 application to Italy with figures published by ISTAT for year 2003. Note that for the application of SM2, the 2003 tax rules have been applied to 1998 ECHP data for the reasons explained earlier.

The table shows the breakdown of total gross income into social insurance, tax and net

components. On the average, net income, after tax and social insurance contributions including employers' contributions, accounts for 70% of total gross.

The table also shows comparison with figures published by ISTAT. *The agreement is reasonable, but not excellent.* Social insurance contributions, in particular employers' contributions, are over-estimated in SM2 compared with the ISTAT figures, while taxes are somewhat under-estimated. Overall, net as proportion of gross income is under-estimated by 2.0 percentage points.

It is expected that the major factor responsible for this discrepancy is that, for reasons explained earlier, *the 2003 tax rules are being applied to 1998 data.* Since there were some significant changes in the tax system between these two dates, some distortion can be expected in the results. Some other substantive aspects of the data also need some closer examination.

Firstly, the situation regarding the employers' social contribution in Italy was more complex than what could be captured from the ECHP data at hand. A proportion of the working population who may have subjectively seen themselves as 'ordinary' employees could in fact be operating under the 'CoCoCo' system. In this later system, both the employer's and the employee's social insurance contributions tend to be much lower than those in the case of 'ordinary' employees. The two categories, however, cannot be distinguished in ECHP data.⁹ Rather than treating all persons reported as employees simply in the 'ordinary' category (subject to high rates of insurance contributions), we used the provisional data provided by ISTAT to estimate the proportion of employees engaged under the CoCoCo system, classified in detail by age, sex and region. These proportions or 'propensities' were used to adjust the estimates of employer's and employee's social insurance contributions at the micro level as follows:

$$S = c_{asr} * S_c + (1 - c_{asr}) * S_o$$

where S_c is the amount of SI contribution under the CoCoCo system, and S_o under the 'ordinary' employment system, and c_{asr} is the estimated proportion among the employees (classified by age-sex-region) who are engaged under the former system.

Table 8 – Application to Italy: comparison with ISTAT

	SM2 (data: ECHP 1998) (tax system: 2003)		ISTAT	Error (% points)
Gross including SI	10,241	100.0	100.0	
SI contributions				
- Employers' contribution	1,361	13.3	11.1	2.2
- Employees' contribution	416	4.1	3.2	0.9
- Self-employment contribution	202	2.0	1.6	0.4
gross taxable	8,261	80.7	84.1	(3.4)
Personal income tax and financial tax	1,044	10.2	11.6	(1.4)
net income	7,217	70.5	72.5	(2.0)
Euro 2003, per capita.				

Sources. ISTAT: National Account (1998). SM2: Italy ECHP Wave 6 Our Model.

⁹ By contrast, it is expected that this information will be available more fully in EU-SILC, judging from the questionnaires used in the EU-SILC pilot in Italy. It distinguishes four different types of self-employment arrangements, including the 'CoCoCo' system, which in reality is more akin to self-employment than to employment.

Table 9 - Distribution of total gross income excluding employers' SI contributions

	SM2 (data: ECHP 1998) (tax system: 2003)		ISTAT	Error (% points)
Gross excluding employers' SI contribution	8,880	100.0	100.0	
- Employees' contribution	416	4.7	3.6	1.1
- Self -employment contribution	202	2.3	1.8	0.5
gross taxable	8,261	93.0	94.6	(1.6)
Personal income tax and financial tax	1,044	11.8	13.0	(1.3)
net income	7,217	81.3	81.6	(0.3)
Euro 2003, per capita.				

Sources. ISTAT: National Account (1998). SM2: Italy ECHP Wave 6 Our Model.

The effect of this refinement is to reduce the employers' SI contribution by around 6%, i.e. from 14% of total gross income, to 13.3% of total gross income reported in Table 8. The estimate still remains somewhat higher than the ISTAT figures. It is interesting to make the comparison after removing the effect of employers' SI contributions, as done in Table 9. As percentage of gross income less employers' SI contributions, mean net income from SM2 is practically identical to the national accounts figures from ISTAT. Some over-estimation in SI components is largely balanced by similar under-estimation of taxes. The results are excellent for the net-to-gross conversion, which are the ones of primary interest for EU-SILC.

In relation estimating taxes, an important point is that deductions and tax credits in Italy depend on information on *household consumption* which is not available in ECHP (nor in EU-SILC) data sets. We have drawn on statistical modelling and matching of data from other surveys to estimate these, and then imputed these on to the ECHP microdata. This requires a considerable amount of work, but there is no other reasonable alternative we think. This is because surveys such as ECHP or EU-SILC, which aim to collect very detailed income data along with a whole range of other variables, cannot be expected also to contain detailed consumption data on the same units.

Even so, the consumption components in the external data used here cannot be distinguished in sufficient detail and some approximation is unavoidable. While the particular statistical procedure we used can certainly be refined, the use of such modelling cannot be avoided for the reasons noted above.

Table 10 shows the distribution of estimated gross income by component. These are the main EU-SILC target variables: all EU-SILC income target variables can be constructed on the basis of appropriate aggregation of such classification by component. The model also provides the same breakdown for net income and the net-to-gross ratio, though these are not included in the required target variables in EU-SILC. Because of differences in component-specific deductions and tax credits, and also in the social insurance contributions, the net/gross ratio varies by component.

The net-to-gross ratio is much lower for income from work (63%) than for other components. This results from the social insurance contributions to which such income is subject. Leaving aside the effect of social insurance contributions, the ratio of *net to gross taxable income* varies approximately from the low of 83% for property income, to 87% for work income (apart from the social insurance deductions, which considerably lower the final net-gross ratios for work income), to 90% for various taxable benefits, to of

course 100% for housing, social assistance and other tax-exempt benefits. These results appear plausible, though external data are not at hand to validate the breakdown in detail by component.

Table 10 – Italy EU-SILC target variables: distribution of income by component

	mean amount		ratio net/gross	% distribution	
	gross	net		gross	net
	(1)	(2)	(3)	(4)	(5)
income from work	7,436	4,691	63.1	72.6	65.0
PY010 employee cash or near cash income	4,319	3,774	87.4	42.2	52.3
employer's SI contribution	1,361			13.3	
employee's SI contribution	416			4.1	
PY050 cash benefits or losses from self-employment	1,137	917	80.7	11.1	12.7
Self-employed SI contribution	202			2.0	
property income	275	228	83.1	2.7	3.2
HY090 interest, dividends, profit from capital investments in unincorporated business	161	134	83.7	1.6	1.9
HY040 income from rental of a property or land	114	94	82.4	1.1	1.3
taxable benefits	2,454	2,221	90.5	24.0	30.8
PY090 unemployment benefits	74	65	88.2	0.7	0.9
PY100 old-age benefits	1,866	1,685	90.3	18.2	23.4
PY110 survivor benefits	316	291	92.0	3.1	4.0
PY120 sickness benefits	3	2	90.2	0.0	0.0
PY130 disability benefits	160	142	89.0	1.6	2.0
HY050 family related allowances	34	34	100.0	0.3	0.5
PY150 other personal benefits	1	1	83.6	0.0	0.0
tax-exemp social transfers	76	76	100.0	0.9	1.2
PY140 education-related allowances	5	5	100.0	0.0	0.1
HY060 social assistance	6	6	100.0	0.1	0.1
HY070 housing allowances	2	2	100.0	0.0	0.0
HY080 regular inter-household cash transfer received	64	64	100.0	0.6	0.9
total	10,241	7,217	70.5	100.0	100.0

Euro 2003, per capita.

7.3 Applications of SM2 to Spain

Table 11 shows the modelled distribution of the total gross income according to social insurance and tax deductions and the resulting net income applied to Spain. According to SM2 results, social insurance contributions account for 17.0% of the total. Income taxes amount to 10.6% of gross income (or 12.8% of gross taxable income). The total net is therefore around 72% of the total gross (meaning 87% of gross taxable). The figures are compared with earlier results published by OECD.

The Table shows a *remarkably good agreement* between the modelled ECHP data and results published by OECD on the overall structure of gross personal income in terms of social insurance deductions, personal income taxes and net income. The OECD results refer to 1997, before the major tax reforms in Spain, while our data are for 1999, after the reforms. Even so, the only significant difference is that, compared to the OECD results, employees' social insurance contributions are a little over-estimated and employers' contributions a little under-estimated by the model.

Table 11 – Application to Spain: comparison with OECD

	SM2 (ECHP1999)		OECD (for 1997)		
	mean amount*	distribution	(1)	(2)	(3)
			Assumed ratio private income /GDP		
			0.67	0.69	0.71
Gross including social insurance	1,377,492	100.0	100.0	100.0	100.0
Employers' contribution	157,072	11.4	12.4	12.0	11.7
Employees' contribution	47,287	3.4	2.8	2.7	2.7
Self-employment contribution	30,450	2.2	2.4	2.3	2.3
gross taxable	1,142,683	83.0	82.4	83.0	83.3
Personal income tax	145,647	10.6	11.0	10.7	10.4
net income	997,040	72.4	71.4	72.2	73.0

*Pesetas 1999 per capita. Sources: OECD: *Revenue Statistics*, Paris (1999). (1)-(3) exemplify sets of figures which can be derived from the published results, depending on the assumed ratio of private income to GDP per capita. SM2: Spain ECHP Wave 7.

Table 12 reports the average gross and net amounts and the net-to-gross ratio. It also shows the distribution of total gross and of total net income by component. All EU-SILC income target variables can be constructed on the basis of appropriate aggregation of such classification by component.

Table 12 – Spain EU-SILC target variables: distribution of income by component

	mean amount		ratio	% distribution	
	gross (1)	net (2)	net/gross (3)	gross (4)	net (5)
income from work	1,096,818	737,057	67.2	79.6	73.9
PY010 employee cash or near cash income	719,829	615,794	85.5	52.3	61.8
<i>employer's SI contribution</i>	157,072			11.4	
<i>employee's SI contribution</i>	47,287			3.4	
PY050 cash benefits or losses from self-employer	142,179	121,263	85.3	10.3	12.2
<i>Self-employed SI contribution</i>	30,450			2.2	
property income	38,948	31,915	81.9	2.8	3.2
HY090 interest, dividends, profit from capital invest	26,659	21,653	81.2	1.9	2.2
HY040 income from rental of a property or land	12,289	10,262	83.5	0.9	1.0
taxable benefits	236,660	223,003	94.2	17.2	22.4
PY090 unemployment benefits	14,011	13,567	96.8	1.0	1.4
PY100 old-age benefits	154,061	143,645	93.2	11.2	14.4
PY110 survivor' benefits	36,406	35,385	97.2	2.6	3.5
PY120 sickness benefits	6,174	6,120	99.1	0.4	0.6
PY130 disability benefits	21,773	20,248	93.0	1.6	2.0
HY050 family related allowances	1,602	1,527	95.3	0.1	0.2
PY150 other personal benefits	2,634	2,510	95.3	0.2	0.3
tax-exemp social transfers	5,066	5,066	100.0	0.4	0.6
PY140 education-related allowances	12	12	100.0	0.0	0.0
HY060 social assistance	115	115	100.0	0.0	0.0
HY070 housing allowances	348	348	100.0	0.0	0.0
HY080 regular inter-household cash transfer receiv	4,591	4,591	100.0	0.3	0.5
total	1,377,492	997,040	72.4	100.0	100.0

[1]: Pesetas 1999 per capita

The ratio of *net* to *gross taxable income* varies approximately from the low of 82% for property income, to 85% for work income, 95% for various taxable benefits, to of course 100% for housing, social assistance and other tax-exempt benefits. These are also the overall net-to-gross ratios for income components not subject to social insurance deductions. These deductions apply only to work income, for which the overall net-to-gross ratios are naturally much lower – at around 67% for employment income and 70% for self-employment income.

These differences are reflected in the resulting structure of gross and net incomes by income component. While gross income from work accounts for nearly 80% of total gross income, it accounts for only 74% when net amounts are considered. By contrast, property income accounts for a somewhat bigger share of the total and benefits (especially tax-exempt benefits) account for a much bigger share of the total when we consider net rather than gross amounts. These results appear plausible, though external data are not at hand to validate the breakdown in detail by component.

8. Imputation and microsimulation

Hitherto we have assumed that data on all relevant income components required for microsimulation are available albeit not in a uniform form. However, there are normally also missing data for which imputations have to be performed. In this section we consider the relationship between microsimulation and imputation.

Imputation refers to the process of using the information existing in a dataset, as well as external information where appropriate, to produce improved estimates for missing, implausible or inconsistent elements in the dataset. Any good micro-level imputation procedure must meet some basic standards: that the imputed values generated preserve the correlation structure of the data, are stochastic rather than deterministic, and are plausible. Here we are concerned with issues beyond such general requirements: special issues involved in the imputation of complex, composite variables. Specifically, the problem is that information on income collected from surveys or similar sources may be reported in different forms by different units and even for different components by the same unit: forms such as net or gross of taxes and other deductions to which incomes are subjects. Such variation in forms must be overcome before normal imputation procedures can be applied. The essential difficulty is the following. Converting the information on income components to a homogeneous form requires *micro-simulation* based on the prevailing tax-benefit system. Generally, such modelling requires that there be no missing values on individual components, even if the available information is not in the required form. That is, imputation should precede modelling. But *imputation* procedures require donor information to be in a homogeneous form, which means that it should be preceded by modelling for the purpose. Below we describe procedures for the application of imputation and modelling routines iteratively and in combination to overcome this problem.

8.1. The imputation procedure

An important characteristic of the income variables is that these form a set in which there is an interdependence between all the components. Hence an appropriate approach is through a multivariate model involving a multiple regression sequence. The sequential multivariate model makes for more complete imputation of the variables, while at the same time safeguarding their variance and their inter-correlation (Raghunathan *et al.*, 2001). Let us consider the following model. Denote with Y the set of k income Imputation

Variables $Y \equiv (y_1, y_2, \dots, y_k)$, with U the basic set of r regressors to be used for imputation, and with $Y^{(t-1)}$ the set of k lagged variables corresponding to the set Y as the supplementary set of k regressors. The set of regressors can be more flexible U_j , including values of the target variables at the preceding wave ('lagged variables') $y_j^{(t-1)}$, and with some variation by variable in the basic set of regressors if required. Clearly, in order not to introduce multi-collinearity, the current and the lagged versions of any of the other variables cannot both be included as regressors. With U_j as the matrix containing variables with no missing data (including as a result previous imputation), and $y_1, y_2 \dots y_k$ are variables with *increasing rates of missing data*, the imputation sequence is determined by the following factorisation:

$$[y_1|U_1] [y_2|U_2, y_1] \dots [y_k|U_k, y_1, \dots, y_{k-1}]$$

where $[Y|X]$ is the conditional joint distribution of Y where X is known. The form of regression depends on the nature of Y , such as a generalised linear regression for continuous variables (as in the case of income amounts), a logistical regression for binary variables, etc.

Application of the procedure requires that (i) all income values have been specified in a standard form, so that issues relating to micro-simulation modelling do not arise, and (ii) variables used as regressors in the model contain no missing values. Missing values are of course encountered in regressor variables. The following systematic approach can deal with this problem.

We have missing values in Y , but we may also have missing values in the regressor sets $Y^{(t-1)}$ and even in U . Now the first step consists in imputing missing values in any variable in order to have full information in the set of regressors. Let $Z \equiv (Y, U, Y^{(t-1)})$ be the entire set of $(2k+r)$ variables considered, and order variables in Z according to the ascending incidence of (proportion of applicable) values missing. Denote with z_i any ordered variable and with Z_i the set of ordered variables from 1 to i . The first set of imputations is defined as follows:

1. Let be j the number of variables in Z with full information.

2. Consider variable z_{j+1} and the set of all preceding variables Z_j . (i) For the imputation of a variables belonging to the subsets Y or U : remove from Z_j all lagged variables for which the corresponding current variable is also present in the set. (ii) For the imputation of a variables belonging to the subset $Y^{(t-1)}$: remove from Z_j all current variables for which the corresponding lagged variable is also present in the set. This gives the reduced regressor set Z_j' .

3. Impute missing values of z_{j+1} using the set Z_j' as regressors.

4. The imputation is done for cases with z_{j+1} missing but known to be non-zero. The donor cases determining the regression are those with z_{j+1} known and non-zero. *The inclusion of zero values in the donor set to impute missing values when they are known to be non-zero can seriously distort the results.*

5. Add imputed variable z_{j+1} to the 'full information' set, and continue the above process till the last (imputable) variable has been imputed.

The objective of this initial imputation is only to complete the set of regressor variables in a reasonable way, and not to provide the 'final' imputations for the target variables Y .

Hence only the imputed values in U and $Y^{(t-1)}$ are retained at the end of this cycle, for use to perform more precise 're-imputations' on Y.

Now we have missing values in Y, but all values in $Y^{(t-1)}$ and U are available or have been imputed at the preceding step. We begin by ordering variables in Y according to the ascending number of (proportion of applicable) values missing; denote with y_i any ordered variable and with Y_i the set of ordered variables up to i.

The first set of imputations:

Let be j the number of variables in Y with full information. Consider variable y_{j+1} and the set of all preceding variables Y_j . For imputing missing values of y_{j+1} , the regressor set Z_{j+1} consists of

$$Z_{j+1} = (U_{j+1} + Y_j + y_{j+1}^{(t-1)}) \quad \left| \quad \begin{array}{l} \text{donors } y_{j+1} > 0, \quad \text{recipients } y_{j+1} \text{ missing but } > 0. \end{array} \right.$$

Here Y_j is the *set* of variables which had none or a lower proportion of missing values (among the applicable cases for each of the variables) than current variable y_{j+1} , and for which any of those missing values have already been imputed. $y_{j+1}^{(t-1)}$ is the lagged variable corresponding to y_{j+1} , and U_{j+1} is the regressor set for y_{j+1} . The imputation is done for cases with y_{j+1} missing but known to be non-zero, and the donor cases are those with y_{j+1} *known and non-zero*.

The requirement to include only the lagged variable corresponding to the current variable being imputed and to confine the donor population to those with *known and non-zero* values on the variable of interest means that the set of regressors has to be varied from one variable to another. This necessitates calling the imputation routine separately for each variable in the sequence. The process is continued till the last (imputable) variable in Y has been imputed.

Table 13 - The first ('triangular') cycle of imputation

Imputation Variable	Donors	Regressor variables	Available or previously imputed variables	Lagged variable
$Y_j = (y_1, y_2, \dots, y_j)$	(Variables with full information)			
y_{j+1}	$y_{j+1} > 0$	U_{j+1}	$Y_j = (y_1, y_2, \dots, y_{j-1}, y_j)$	$y_{j+1}^{(t-1)}$
y_{j+2}	$y_{j+2} > 0$	U_{j+2}	$Y_{j+1} = (y_1, y_2, \dots, y_{j-1}, y_j, y_{j+1})$	$y_{j+2}^{(t-1)}$
.....			
y_k	$y_k > 0$	U_k	$Y_{k-1} = (y_1, y_2, \dots, y_{j-1}, y_j, y_{j+1}, \dots, y_{k-1})$	$y_k^{(t-1)}$

Imputation cycle for the full set:

After all variables in Y have been imputed once, the following cycle is applied iteratively. The variables are ordered as before, in increasing proportion of the originally missing values. In each application, the full set of regressor variables is used, using all values, whether originally available or imputed in previous cycles. As before, the imputation routine is called separately for each variable in the sequence. The sequence is repeated a number of times. Based on our experience with ECHP data, a few (3-5, say) cycles should be sufficient in most cases.

8.2 Imputation and Modelling in Conjunction

In order to appreciate the interaction between imputation and modelling systems, consider matrix Y , the set of imputation/modelling variables for a set of units.¹⁰ We denote by y_i a particular variable in the set, or where necessary, by y_{ji} that variable for unit j in the data set. In any cell (j,i) of this matrix, the value of the variable may appear as or easily transformed to the form $[0, N_i, H_i, X]$, i.e., the value is either zero (no income from the source concerned), missing (X), specified as the net amount (N_i), or in one of the various possible forms which can be converted to the form H_i even in the presence of missing values on other variables for the unit.

Net-Gross modelling applies along individual rows of the matrix Y , to one unit at a time. However, in order to 'model' i.e. construct the full information on gross and net amounts for each component, it is necessary that there be no missing values (X) in any cell of the row for the unit (that is, as noted, conversion between N_i and H_i generally requires that there are no such missing values). Hence modelling requires prior imputation of the missing values.

Though imputation for missing values can be carried out along columns of the matrix, i.e. variable by variable, it invokes the whole matrix in order to take into account the correlation between variables and between units. It is necessary that for each variable, the available information for all units is in the same form (always net, always gross etc.). The form may vary from one variable to another, but should be uniform for all units within each variable. Where this is not the case in the data as collected, imputation requires prior modelling to meet this requirement.

Hence the combined imputation and modelling system has to be interactive and iterative. The whole process is described step by step below. **Step (0)** provides the starting point by using the micro-simulation model to rationalise the input information, and provides some initial information on the basis of 'complete information cases', i.e. units with no missing data. **Step (1)** concerns the conversion of the collected data for each variable into a uniform form for all responding units. It is only on this basis that imputations for missing values for the variable can be performed. On the basis of these initial imputations for the current substantive variables (income components), imputations for lagged versions of these variables and for regressor variables are performed in **Step (2)**. Our primary concern is with *current substantive variables*. Hence the results for the regressor and lagged versions from Step (2) can be considered as 'final', without the need for further iterative refinement. **Step (3)** produces refined imputations for the substantive variables, using the regressors and lagged variables imputed earlier. The procedure is applied iteratively.

Step (0). Initial data conversion and modelling

0.1 Variables (i) are ordered according increasing proportion of missing values among applicable cases.

0.2 Units in Y are divided into two subsets:

A: complete information set (units with information available on all variables);

B: set with missing values on one or more variables.

0.3 The starting point of this process is provided by the 'complete information' set of units (set A), for which there are no missing values on any variable and hence modelling

¹⁰ Imputation and modelling will generally involve different sets of variables. For simplicity and clarity, it is assumed here that the two systems involve an identical set of variables.

can be carried out without involving imputation, and the data for set A reduced to the following form, giving amount both in gross taxable (H_i) and in net (N_i) forms for each income component (i), and also the unit's 'tax rate' R : $[0, H_i, \text{and}, N_i, \text{and}, R]_{j \in A}$.

0.4 Using the model conversion routine, the available information for units in set B is reduced to the form: $[0, N_i, H_i, X]_{j \in B}$.

Step (1). Conversion to uniform form

1.1 For each variable (column), the predominating reporting form (Y_i) is determined. This is done on the basis of whether H_i or N_i is the more common form of reporting among units in set B:

if $\text{count}(H_i)_{j \in B} > \text{count}(N_i)_{j \in B}$ **then** $Y_i \equiv H_i$ **else** $Y_i \equiv N_i$

1.2 As a starting point, we may assign *the average* of R_j values for units in set A (say, R_0) to every unit in set B. This permits the conversion of all reported values (N_i or H_i) for each variable to its 'predominating form' Y_i defined above using the modelling relationships. For set A the information in any cell is already available in the 'predominating form' from Step(0)¹¹: Set A $\rightarrow [0, Y_i]$. For set B the resulting form is: Set B $\rightarrow [0, Y_i, (Y_i | R_0), X]$.

Here, form ' Y_i ' indicates the original information was already in the 'predominating form' Y_i , so that no transformation using R is required. Form ' $(Y_i | R_0)$ ' indicates that the original information was specified in the form different from the predominating form Y_i for the variable, so that transformation to Y_i conditional on the assumed R value was required.

1.3 Now consider the parallel set of lagged variables $Y^{(t-1)}$. The reporting forms can be simplified as above. Then rows and columns of this matrix are arranged identically to their arrangement in table for Y , units and variables in the same order as Y . Also, the R_j values and the *form* Y_i from matrix Y are imposed on $Y^{(t-1)}$. (The actual values are of course different; it is the choice between N_i and H_i forms which is imposed from t on $t-1$.) These assigned parameters are then used to transform existing $Y^{(t-1)}$ cell values into the same form as that for Y in step (1.2) above. The resulting form is: cell values of $Y^{(t-1)} \rightarrow [0, Y_i^{(t-1)}, (Y_i^{(t-1)} | R_j^{(t)})], X]$.

Here (t) indicates that the R values are taken (copied) from the current (t) data set, as is the predominating reporting form Y_i . It has been considered preferable to borrow the arrangement, parameters and data forms from the current set Y , since it is not generally appropriate to apply the tax model for the current year to past data.

Step (2). Imputation of regressor and lagged variables

2.1 The resulting information is now in a form such that the imputation procedure above can be applied exactly as described to produce a preliminary complete set of all current (Y), lagged ($Y^{(t-1)}$) and other auxiliary variables (U).

2.2 The above is still based on preliminary estimates of R -values for the less than complete information subset B. With missing values already removed, we apply the model to produce improved estimates of R_j for all units (Table 14).

¹¹ While the predominant form is determined from set B only, data are converted into this form for both sets A and B.

Table 14 - Data form after Step (1.2)

		Variable						Tax	
Unit		1	2	...	i	...	I	rate	
SET A	1							R_1	
	
	j							$[0, Y_i]$	R_j
	a								R_a
SET B	a+1							R_0	
	...							$[0, Y_i, (Y_i R_0), X]$	R_0
	J								R_0
		Y_1	Y_2	...	$Y_i (=H_i \text{ or } N_i)$...	Y_I		

predominating reporting form (determined by set B only)

2.3 These improved values of R_j are imposed on corresponding rows of $Y^{(t-1)}$.

2.4 All values imputed at (2.1) above are rejected, from all sets Y , $Y^{(t-1)}$ and U .

2.5 The remaining original values in both Y and $Y^{(t-1)}$ are transformed to the predominating form Y_i using the R_j values as defined in (2.2) and (2.3) above.

2.6 The imputation procedure at (2.1) is now repeated on the resulting data set. These are taken to be the final results for the lagged and regressor variable sets.

2.7 For the main data set Y , the results from (2.6) are used to re-estimate R_j values, but the imputed values of y_i themselves are rejected. The remaining original values are transformed to the predominating form Y_i using these improved R_j values.

Step (3). Imputation of target variables Y

The data form is now the same as that in Step(1) for set Y . The only difference is that, where necessary, all of the available values have been converted to the predominating form Y_i for variable i conditional on current values of parameter R_j specific to each case j , the cell values being: $[0, Y_i, (Y_i | R_j), X]$.

The first ('triangular') imputation

3.1 The first ('triangular') cycle of imputation is performed as in Table 13, without distinguishing between the two forms of known values Y_i , giving cell values in the form $[0, Y_i, (Y_i | R_j), Y_i']$, where prime (') indicates imputed values.

3.2 The resulting complete set is used, with the SM2 procedures to obtain improved values of R_j for each unit.

3.3 Finally, these R_j values are used to re-estimate the conditional values $(Y_i|R_j)$.

Imputation cycle for the full set

3.4 Next is performed imputation cycle for the full set using SM2. As explained, imputations (3.1) and (3.4) are performed on variables in order of increasing proportion of

missing values, but using imputed values for all other variable previously imputed in any cycle.

3.5 The resulting complete set is used with the modelling procedure to obtain improved values of R_j for each unit.

3.6 These R_j values are used to re-estimate the conditional values ($Y_i|R_j$).

3.7 The sequence is repeated a number of times.

9. Concluding remarks

The Siena Micro-Simulation Model (SM2) has been developed as a practical tool for the purpose of providing a robust and convenient procedure for the conversion between net and gross forms of household income. Its major attraction is the ability of the standardised core procedures to handle diverse tax regimes. The context of its implementation is international: specifically the survey data on income and living conditions collected in EU countries. In these respects, SM2 differs from most existing micro-simulation models and procedures.

The usefulness of this tool depends on how widely it is used, both by official statisticians and by social researchers generally. To facilitate such use, University of Siena makes available freely the SAS programs and related documentation for the procedure. These resources are also available from Eurostat who have officially adopted the SM2 procedure in the context of the EU-SILC programme.

References

- Betti, G., Verma, V., Ballini, F., Natilli, M. and Galgani S. (2003), Statistical imputation in conjunction with micro-simulation of income data, *Rivista Italiana di Economia, Demografia e Statistica*, **57(3/4)**, pp. 35-44.
- Euromod (2001), *Euromod: an integrated European benefits-tax model*, Final Report, edited by Sutherland, H., Euromod Working Paper n.EM9/01.
- Eurostat (2002), Commission Regulation concerning EU-SILC as regards the list of primary target variables, EU-SILC107.
- Eurostat (2004), *Income in EU-SILC: Net/Gross/Net conversion. Report on common structure of the model; model description and application to the ECHP data for France, Italy and Spain*, prepared by V. Verma, G. Betti and co-researcher. EU-SILC 133/04, Luxembourg.
- Raghunathan T. E., Lepkowski J., Van Voewyk J. and Solenberger P. (2001) A Multivariate Technique for Imputing Missing Values Using a Sequence of Regression Models, *Survey Methodology*, **27**, pp. 85-95.
- Verma, V., and Clémenceau, A. (1996), Methodology of the European Community Household Panel, *Statistics in Transition*, **2(7)**, pp. 1023-1062.